Hephaestus: Data Reuse for Accelerating Scientific Discovery

Jennie Duggan
Northwestern EECS
jennie.duggan@northwestern.edu

Michael L. Brodie
MIT CSAIL
mlbrodie@csail.mit.edu

ABSTRACT

Data-intensive science, wherein domain experts use big data analytics in the course of their research, is becoming increasingly common in the physical and social sciences. Moreover, data *reuse* is becoming the new normal, owing to the open data movement [15] and arrival of big science experiments such as the Large Hadron Collider. Here, a small group of researchers with exotic equipment produce a dataset that is shared by thousands. Unfortunately, weak and spurious correlations are also on the rise in research [5, 27]. For example, Google Flu Trends published their algorithms in 2008 [19] for use in public health, and in the intervening time its accuracy has plummeted. In the 2011-2012 flu season, this system produced estimates more than 50% higher than the number of cases reported by the U.S. Center for Disease Control [32].

This work first examines common pitfalls associated with data-intensive science and how they contribute to irreproducible results. We then propose a system for conducting virtual experiments over existing data. It simulates randomized controlled trials by reframing the principles of empirical research. These virtual experiments underpin a larger platform we call Hephaestus. This framework accumulates virtual experiments in a visualization to help scientists identify consistencies and anomalies in an area of research. We then highlight a set of research challenges associated with this platform. We argue that by using this approach, data-intensive science may come to achieve accuracy on par with its causality-driven predecessors.

1. INTRODUCTION

Data reuse is becoming increasingly prevalent in science. The reasons for this are numerous. First, many science funding agencies are instituting open science mandates [17, 41], and this promises to create a new flood of data sourced from published research. In addition, many large-scale science endeavors are now designed to collect data *first* for use in any number of studies later. For example, the Large Hadron Collider makes its measurements available to more than 8,000 scientists, although relatively few people operate the particle accelerator [10]. This pattern is echoed in the Large Synoptic Survey Telescope [52], the Square Kilometer Array [9], NASA's MODIS satellite imagery [38], and many others.

There are many challenges and opportunities associated with making use of this growing body of data. This data reuse calls for new techniques because it changes how scientists conduct their research. Traditionally researchers begin with a hypothesis followed by an experiment designed to prove or disprove it. These trials are conducted in carefully controlled lab settings. In each experiment, the researchers manipulate perhaps one or a few variables to establish cause-and-effect relationships. It is unclear how these rigorous, time-tested methods will evolve for scientific discovery over existing data.

To explore these questions, we interviewed scientists to learn about how they conduct data-intensive research now. Their experiences span a variety of empirical disciplines, including evolutionary biology, genomics, clinical trials, and photonics. Our interviews revealed a widespread need to reframe the principles of empirical research for eScience. These discussions also brought to light numerous pitfalls in data-centric analysis; we discuss some of these hazards in the context of machine learning below.

Machine learning alone is not enough Researchers presently use statistics and machine learning to discover interesting correlations from their experimental results. There has been considerable excitement about this development with many heralding it as the "end of theory-driven science" [3, 24]. Recent evidence, however, suggests that one quarter of research is statistically false [28], and others estimate a much higher rate [5, 27]. Although some of these errors may be attributed to shoddy research procedures or buggy code, the prevailing wisdom is that such failures are owing to both the limits to and the misapplication of statistics over massive datasets [12, 22, 45].

A plethora of examples illustrate this issue. Google Flu Trends, in conjunction with the United States CDC, published statistical models in Nature for use in predicting seasonal illness rates [19]. Their approach used search engine queries to predict the rate of people seeking treatment for influenza-like illnesses. This technique was pitched as an early detection method for flu pandemics. As time progressed, it became clear that Google Flu Trends was vulnerable to overfitting; its error rates skyrocketed in subsequent years [35]. In particular, as the tool became more well-known, users queried it at a higher rate, confounding its results.

In unrelated research, epidemiologists published observational studies demonstrating a positive correlation between post-menopausal women taking hormone replacement ther-

7th Biennial Conference on Innovative Data Systems Research (CIDR '15)
January 4-7, 2015, Asilomar, California, USA
.

apy (HRT) and a reduction in heart disease [21]. This result puzzled many experts in the field, because there is no intuitive link between the two. Despite this skepticism, the finding was used to promote HRT to this demographic. Later, the link was repeatedly refuted with randomized controlled trials, reversing this policy recommendation [31]. This error was attributed to the initial study's use of subjects from a single socioeconomic group, an unrelated variable that was not controlled for in the study's design.

Earthquake modeling has also demonstrated some high-profile failures in its predictions. In 2006, seismologists predicted that an area in the Indian Ocean was at low risk. In September of the following year, an 8.5 magnitude event struck at exactly that location, discrediting this analysis. [49]. To this day, researchers struggle with this issue, but their success has been stymied by an inability to measure the underlying causes of the quakes. Scientists can measure an event only when it is occurring.

Machine learning and statistics have shown immense use in tackling real-world problems, such as pattern recognition for manufacturing defects and expert recommender systems. The needs of science—in hypothesis creation and testing—are fundamentally different from the aims of machine learning. Whereas the latter looks for actionable patterns in data, it does not speak to the root causes of an outcome. In contrast, the scientific method uses carefully designed experiments to test for cause-and-effect relationships. Machine learning also differs from the statistics used in science because its transformations and results rarely use error bars denoting the quality of their predictions. In empirical science, results always come with error bounds.

What changes with data reuse? We submit that as science data becomes plentiful, it will dramatically alter how research is conducted. The eScientist's principle artifact or work product will be the experiment that she will design, develop, incrementally test, validate, and publish directly on top of massive data sets. For this, one will need a language and data management platform.

Eliminating the current data acquisition bottleneck frees eScientists to focus on their primary contributions, namely accurately modeling data that represents phenomena. More specifically, *how* a researcher formulates a problem, and whether that conception holds up to observations collected from many sources, will become more important than *who* collected the data. Hence, enabling eScientists to directly express, manipulate, and test collections of hypotheses is needed. In this context, models are analogous to rulesets that express causal relationships. Indeed, it will be crucial to automatically identify the conditions under which the data agrees with the model and when the two diverge.

As data accumulates from many disparate sources, it will become too large for researchers to download and query on their own. Also, it is unlikely that the data they want to analyze will be all located on a single host. Hence, eScientists will need a means of creating queries and orchestrating their execution to test hypotheses. This new challenge is at the intersection of data management and statistics.

In addition, open science data will make it possible to subject discoveries to continuous verification. As new data arrives, especially from studies that build upon the prior work, people further test their findings. Over time, this will enable researchers to distinguish short-term correlations from long-term cause-and-effect relationships.

It is likely that these properties of data reuse will incentivize ease of use and transparency in its application. When it is clear how the research associated with a publication was conducted, weak and spurious correlations will be more readily identifiable. Ideally, this will increase the accuracy of follow-on work. Right now, most efforts in this area focus on workflow management [16], but we argue that this approach attempts to make programmers of scientists. There are many tools for processing raw measurements into data products [8, 25]. In this work we focus on the analysis of data that has already been cleaned and labeled.

Man-Machine Symbiosis It is our position that data science tools should *augment* the capabilities of human researchers rather than supplant them. Computers alone lack the deep domain knowledge needed to semantically break down the space of possible hypotheses into tractable subproblems, and it is not clear that a solution to this issue is on the horizon. On the other hand, humans are only capable of reasoning about models of limited complexity, with fewer than ten concepts in short-term memory [37]. To realize the complementary strengths of empirical research and data-intensive science, we embrace man-machine symbiosis in the tradition of Licklider [33]. Rather than mining the data, it is our goal to help scientists *search for cause-and-effect relationships.*

Our approach focuses on human-guided exploratory analysis rather than deferring to automated scientific discovery as in [47, 48]. A recent survey of open problems in data mining concluded that human verification of machine-discovered relationships will be needed for the foreseeable future [12]. Hence, it is important to carefully consider the structure of this partnership.

Virtualizing the Scientific Method Randomized controlled trials are the gold standard for proving causality in many domains of science. The central building block of our vision is the *virtual experiment* (VE), a hypothetical language with which scientists would design, develop, test, execute, and publish data-intensive research. VEs are part of a larger platform that we call *Hephaestus*[1], a meta-system that enables users to create and execute experiments over local and remote big science data. We call it a meta-system because it sits on top of existing science databases that execute complex analytics locally. Using correlations that are verified by experts, Hephaestus will assemble *probabilistic causal graphs*, as defined in Section 3.2.

VEs will empower researchers to focus on *experimental design*, abstracting away the underlying plumbing, e.g., where the data comes from and how the query will run. This experimental design taps in to any number of data sources, which may be stored locally or remotely. Working at this level will enable researchers to focus on exploring the space of possible theories working hand-in-hand with the meta-system.

VEs will also enable scientists to report their research protocols in a standardized fashion. Hence, when one publishes using grant money tied to an open data mandate, she may send her results to an open science repository with VEs for reproducing her experiments. More importantly, these VEs will let others understand and expand on these results thus contributing to the fundamental objective of open data —accelerating scientific discovery. The U.S. National Institutes of Health has created two initiatives for this goal [39, 40].

[1] Named after the toolmaker of the gods of Olympus who built automatons of metal to work for him.

Naturally, Hephaestus will aid the reproducibility efforts of scientists by enabling them to compose and share hypotheses. Also, if their new experiments build on prior findings, the scientist can verify that their assumptions are correct. The data management community is clearly in a position to help solve this challenge.

This study extends the rich and challenging research area of computational platforms for data-intensive analysis [42, 30, 50]. Our focus, however, is on data reuse for eScience and on reframing concepts from empirical research, whereas their approaches are more closely aligned with machine learning and knowledge discovery. This proposal is distinct from an electronic lab notebook (ELN) [46]. It is designed for probing massive hypothesis spaces rather than improving data processing workflows and maintaining provenance for specific data.

This rest of this paper is organized as follows. In Section 2 we briefly summarize the principles of empirical research and how we formulate the challenge of data reuse. Section 3 outlines our vision for the Hephaestus meta-system. In Section 4, we delve into the open challenges associated with this work and conclude.

2. BACKGROUND

It is our goal to lay the foundation for extending concepts from empirical science to data reuse so that researchers can directly and declaratively design experiments over massive, open datasets. In this section, we briefly touch on the some of the terms and methods most relevant to this study. We then discuss how Hephaestus fits into the context of the current practices in scientific research.

2.1 Principles of Experimental Design

Scientists search for causal relationships. They do so by making predictions that are readily falsifiable, or capable of being disproven. These relationships describe when an *intervention*, or measurable action, creates an *effect*, the outcome that is under prediction. [44] Whereas machine learning seeks out correlations with strong *predictive* power, scientists pursue ones with strong *explanatory* power, a subtle but important distinction. For a relationship to be causal, it must also have *validity*, such that it generalizes to previously untested circumstances covered in the initial theory.

Statistical hypothesis testing is a long-standing convention in empirical research, especially in the social sciences. Economics is one such discipline. When analyzing a large, dynamic system like a country's economy, the scientist's only option is to obtain data collected previously without specific controls. Conducting experiments by applying an intervention in a controlled setting is not possible for them. Hence, they use well-developed statistical tests to evaluate their theories. VEs will confirm or deny hypotheses using the same techniques for data reuse. This will make it possible to scalably test many hypotheses, because the system will be able to rapidly rule out many of them automatically. Therefore, after the scientist has designed their experiment, they consider only those correlations that pass the test, rather than manually wading through a barrage of superfluous ones.

Researchers use statistical hypothesis testing to determine whether a result is *statistically significant* or unlikely to have occurred by chance alone. Starting with a hypothesis, the experiment designer selects a *null hypothesis* that defines the anticipated experiment outcome if the intervention has no effect. For example, in clinical trials for new drugs the null hypothesis is usually quantified using a placebo group. They then propose an *alternative hypothesis*, or result if the theory under test is correct. The clinical trial would use the data from patients given an experimental drug to test the alternative hypothesis.

Once these two competing hypotheses are established, the scientist decides how to compare them, frequently with a p-value, although other metrics are also used. The p-value is used to reject the null hypothesis by calculating the probability that the outcome observed in the presence of the intervention would have happened by chance alone. This figure needs to be below a threshold for the theory to be judged successful, and most disciplines use a threshold of 0.05. A test with a p-value of 0.05 implies that the null hypothesis has a 5% chance of being true. This threshold, and how to select it, has been the subject of intense debate in recent years. Hence, a statistical hypothesis testing framework needs to be sensitive to evolving standards for testing a causal link.

Causal relationships improve upon the null hypothesis. Correctly designed controls are an important staple of nearly all empirical studies. In practice, we found that most data-intensive science uses one of three types of controls. A *sampled* null hypothesis measures the experiment's conditions in the absence of an intervention, as in the clinical trial example above. A *synthetic* control is a constant or probability density function supplied by the user. For example, when researchers at the Large Hadron Collider were searching for the Higgs Boson, they used p-values for hypothesis testing. Because they were trying to determine whether or not the particle exists, they had no way of measuring the absence of a discovery. Hence, they used a probability density function to describe background noise for their control. Sometimes controls are formulated as tests of *independence*, where the null hypothesis presumes that no relationship exists between two or more variables. . A researcher asking whether there is a statistically significant link between gender and heart disease might use this type of control.

Naturally, Hephaestus will need to support all three of these approaches to control design. Clearly, picking the right one for a given theory is not simple, and domain expertise will be critical for this part of the VE design.

Pitfalls There are several challenges that arise when researchers use statistical hypothesis testing, and here we list a couple of prominent ones. Test designers need to be vigilant about *confounders* or extraneous interventions that are covariant with the target effect. It has been reported that from 1998 to 2007 the diagnosis rate of autism was strongly correlated with sales of organic food [36]. Although these two variables are correlated, and this might pass a statistical hypothesis test, organic produce is a confounder for this disorder. There are some techniques for detecting the existence confounders at small scale [44], but they cannot determine the variable is responsible for a spurious correlation.

Another issue that comes up with statistical hypothesis testing is *lurking variables.* These variables have an effect on the experiment outcome, but are not included in the analysis. Simpson's Paradox is one instance of this pitfall. Here, a trend that is present in data that is binned into groups disappears or is reversed when the data is aggregated. We illustrate this issue with an example from a study of kidney stone treatments [11]. The authors compare the efficacy of two treatment options, A and B. They first consider two patient populations, one having small kidney stones, and the other having large ones. For the small group, A is effective for 93% (81/87) of patients, and B works for 87% (234/270) of them. The second population had a success rate of 73% (192/263) for A and 69% (55/80) for B. It would appear from these results that Treatment A is the clear winner. On the other hand, if we combine the groups, A cures patients at a rate of 78% (273/350) and B helps 83% (289/350) of the time. Here, Treatment B appears to be the best choice. As we will see in Section 3.1, it would be easy to make a VE for either scenario, and human intervention is needed to select the right course of action. Both of these issues pose greater challenges over massive datasets, where the number of variables and complexity of the interactions rises.

2.2 Data Reuse Goals

In addition the principles of empiricism, there are several other factors that shape the needs of scientists in the context of data reuse. Every discipline has agreed-upon standards for how they test hypotheses statistically. In addition, their inquiries may take the form of incremental steps or big picture inquiries. Their interactions with the data are very different when researchers are analyzing anomalies as opposed to confirming existing theories.

Community Standards Practically every scientific discipline has community standards that dictate how its practitioners apply statistical hypothesis testing to their discoveries. Each community has preferred methods for constructing controls, comparing them against the alternative hypothesis, and thresholds of significance. These practices are used in peer review to confirm or reject a new discovery.

For a VE platform to aid in testing of new theories, it needs access to libraries that capture these best practices. Naturally, these libraries need to be extensible to follow the norms of a community and support new techniques as they arise. Presently, we are seeing this evolution happen in the life sciences, where researchers are starting to adopt bayesian hypothesis testing in lieu of frequentist approaches [28].

Discovery Approaches To identify the high-level tools needed for data reuse, we draw from Kuhn's study of the history of scientific discovery [29]. In it, he argued that research happens in two flavors: normal science and occasional periods of revolutionary science. *Normal science* works within a *paradigm*, building on a set of accepted discoveries that provide a coherent "model of the world" for follow-on work. This research is incremental and discoveries of this kind are usually predictable by practitioners of a field. Take for example Boyle's Law, which codifies the relationship between gas pressure and volume. In Kuhn's taxonomy it is considered normal science because it built on established theories of thermodynamics. Although this law is still in use today, Boyle needed this paradigm to exist before he knew the right questions to ask. The majority of science uses this "puzzle solving" approach to discovery, and we designed VEs to support these questions.

Over time, normal science may accumulate data that exposes the limitations of a paradigm. If these anomalies are consistent—implying that certain parts of the paradigm are incorrect or incomplete—then this opens the door for rival frameworks. Kuhn terms these shifts "revolutionary science" because they challenge long-held and seemingly obvious assumptions. When Copernicus theorized that the earth revolves around the sun and not the other way around, this was a new paradigm. Initially, his theory did not work using existing tools for calculating planetary motion and new methods were needed to make accurate predictions about the location of celestial bodies at a given time. To this end, we also need to create tools so that scientists can evaluate how individual contributions, such as a single publication, fit into the larger context of their field. We propose probabilistic causal graphs in Section 3.2 to help scientists probe the strengths and limitations of the paradigms within which they work.

In summary, we design the two main components of Hephaestus to address the ways that scientists conduct their work. VEs will be useful for theory-driven normal science that investigates discoveries in the context of a larger paradigm. Here, the scientist proposes a relationship for study that confirms an existing system of rules and standards. On the other hand, probabilistic causal graphs will help scientists examine the broader implications of their work by assembling collections of discoveries so that researchers can relate them back to the underlying assumptions of their experimental design. This data-driven strategy will help scientists look for results in the data that consistently contradict the state of the art.

3. HEPHAESTUS

We now take a look at our proposed eScience open data platform, Hephaestus. It consists of two parts: virtual experiments and probabilistic causal graphs. The former is designed for exploring relationships pertaining to a small number of variables. VEs will do so by executing statistical hypothesis testing over existing data. This analysis is well-suited for identifying causal links to a specific phenomenon. On the other hand, probabilistic causal graphs will target scientists looking at their research at a high level. These graphs maintain a large number of relationship derived from VEs so that researchers can explore paths of proposed causality identifying consistencies and anomalies in a body of work. This approach will help scientists evaluate their work holistically and is amenable to comparing competing scientific paradigms.

3.1 Virtual Experiments

The scientific method enables experimenters to produce empirical data for a specific experiment. As we saw in the previous section, designing an experiment is not a trivial undertaking, and doing it correctly is crucial for meaningful results. Recall that VEs are designed to simulate randomized controlled trials. Below, we outline the requirements of these trials, and how they might in principle translate to VEs. Randomized controlled trials call for:

- **Controls**: Trials contain a test condition and a control to verify that the target effect only occurs when the intervention is applied. If VEs are conducted over the results from published studies, in many circumstances they will reuse existing controls. Sometimes, however, the control needs to be either estimated with a model or calculated from other sources, as discussed in Section 2.1.

- **Blocking**: Lab experiments also include blocking, where samples are divided into disjoint sets to evaluate the hypothesis separately over naturally occurring sources of variance. In the parlance of experimental design, blocking is often expressed as “controlling for x, y, and z”. VEs will take in blocking parameters describing how the data will be partitioned for evaluation. Blocking also dictates how samples are selected for reuse, by specifying traits that need representation. Getting an experiment’s blocking correct is crucial for avoiding lurking variables.

- **Randomization**: Researchers assign subjects to groups (control or test) without explicit selection. When a VE reuses data, samples have either already received the intervention or did not. The new trial defers to the randomization applied in the initial experiments. In some VEs, such as chemistry experiments, randomization is not necessary.

- **Repeatability**: Empirical trials replicate their findings, collecting enough samples to identify natural sources of variation and to have high statistical power. VEs may have minimums on the sample sizes needed to complete their calculations, such as being a representative proportion of a known population.

Running Example We now introduce a running example in a strawman query language to motivate this work. This example works within a single relational-style engine, but it is easy to envision extending it to any number of data models and storage engines.

A VE evaluates one or more hypotheses to either identify the most promising ones or to determine conditions where human judgment is needed, e.g., to resolve confounders. VEs start with an effect under study and may suggest one or more interventions. If an oncologist is studying the root causes of skin cancer, she might begin with hypotheses of her own, such as sun exposure and fair skin. She would pose the query:

```
SELECT * LIMIT 10
FROM  cancerSubjects
EFFECT 'skin cancer' as S
INTERVENTION  sun exposure, skin tone, *
ANALYSIS count(S)/count(*) as c
CONTROLLING FOR age, gender
SCORE BY pvalue(c)  ASCENDING
WHERE  pvalue(c) <= 0.05;
```

It returns a set of hypotheses, potentially including the interventions listed, ranked by a user-supplied scoring function that estimates each’s likelihood of a causal relationship if evaluated in a lab under rigorously controlled conditions.

3.1.1 Experiment Definition

The VE’s parts are:

- **Sources** The FROM clause designates a single source of data, 'cancerSubjects'. In implementation, VEs may draw from numerous data sources, and individual ones may be local or remote. Hephaestus translates its queries into the language of the underlying database in an external compilation step.

- **Interventions** It has one or more interventions for evaluation. The wildcard, ‘*’, denotes the asker’s desire to discover additional hypotheses using Hephaestus. This initiates a search for interventions that will pass the user’s hypothesis test.

- **Effect** The EFFECT keyword denotes the outcome under study. It is a value or range of values for an attribute in the source data.

- **Analysis** Within each blocking group’s control and test sets, the VE performs analysis to summarize the success of the intervention or characterize a correlation. The oncologist calculates the percentage of samples testing positive for skin cancer, where the effect is a binary value. The outcome of analysis is passed on to the scoring function.

- **Controls** A VE’s blocking is declared with CONTROLLING FOR. If the cancer subjects are binned by age into k discrete groups, and it has two values for gender, this experiment runs $2k$ blocks per hypothesis, each having a control and test group.

- **Scoring Function** The user supplies a function for Hephaestus to use when ranking each hypothesis. It consists of one or more *accuracy metrics*, The oncologist scores a hypothesis by how it stacks up against the null hypothesis with p-values.

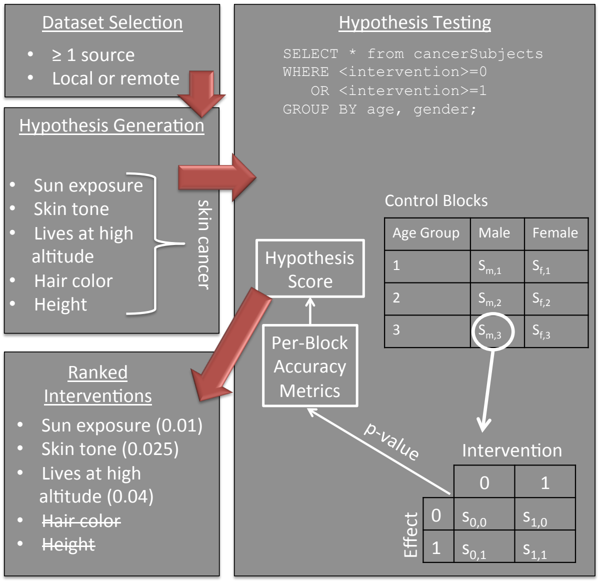
- **Limit** Although the system may evaluate any number of correlations, the experimenter is free to limit the number returned in order to expedite the query and avoid burying the researcher in results. This is a generalization of top-k querying [26].

- **Thresholds** Recall that many disciplines only accept hypotheses having statistical significance over a threshold in their peer-reviewed literature. The VE captures this using the WHERE clause, and this denotes that the query should not rank interventions below this bar.

- **Uncertainty** The source data for VE experiments may contain uncertainty owing to human error or sensor imprecision. Query writers can also inject uncertainty if they are not confident about their source data or the applicability of their analysis techniques, reflecting their domain expertise. The VE executor will need to incorporate this into its analysis. It is unclear whether the best approach for this is many worlds modeling [14], probability density functions [18], or ranges [54]. The model may be dependent on the application. In addition, the experiment may produce error bars in its results; these are most commonly confidence intervals or standard errors.

3.1.2 Query Execution

Figure 1 has an overview of the logical steps of a VE from the user’s perspective with the running example. The query begins with a dataset selection, and this consists of one or more sources residing in any number of external databases. The VE continues by verifying the supplied hypotheses against the schema and searching for new ones to satisfy the wildcard clause. It then scores each hypothesis, either in parallel or sequentially. The per-hypothesis scoring starts by assigning all of the samples with a recorded effect of skin cancer or the absence thereof to a control block such as ”female-1”. The VE evaluates each block, subdividing its samples into control and test groups. It then performs the analysis, taking the percent of samples affected by cancer in each group for the p-value and aggregates over the blocks



```
SELECT * from cancerSubjects
WHERE <intervention>=0
   OR <intervention>=1
GROUP BY age, gender;
```

Age Group	Male	Female
1	$S_{m,1}$	$S_{f,1}$
2	$S_{m,2}$	$S_{f,2}$
3	$S_{m,3}$	$S_{f,3}$

	0	1
0	$s_{0,0}$	$s_{1,0}$
1	$s_{0,1}$	$s_{1,1}$

Figure 1: Virtual experiment logical model.

for each hypothesis. The query concludes with a set of interventions, ordered by the relevant scoring function. Three interventions meet the p-value threshold, and they are ranked by their statistical significance.

Dataset Selection The first step of the VE is dataset selection. Once Hephaestus has the VE's schema, it can compose queries to the source database(s) for hypothesis testing. To get started, the VE will rely on the user to declare a VE's sources, matching their schemas with the experiment's parameters, including the effect and interventions. In future work, Hephaestus would benefit from search functionality to aid users in locating datasets relevant to their queries. This search will need to leverage the metadata associated with its sources to determine the VEs matches In the long run, ontologies and metadata will be key to enabling researchers to find data sources for their VEs; they may enable the engine to automatically infer input schema mappings to clauses in the VE query. One could imagine a search engine for data that takes in a string describing the correlations the user would like to explore and returns a ranked list of potential sources for the query.

Hephaestus will need ways to determine how suitable different datasets are for a given VE. This will be a function of how well the dataset corresponds to the controls, anticipated distribution of samples, et cetera. The engine can estimate this by tapping into metadata about how a dataset was collected and processed.

If a dataset corresponds perfectly with the VE's design, we say it is *empirical* in this context. If a scientist is reporting the results of a published study, and they submit a VE accompanied by the data collected for the paper, this is empirical data. It is effectively the same as a conventional experiment. On the other hand, if a dataset comes with limited metadata about its provenance and contents, we say that it is *abstract*. The quality of a candidate source will vary between these two endpoints and its value on this scale will also depend on the VE.

It is an open question for eScience practitioners and statisticians what level of metadata matching is necessary to prove causality. As open data becomes more prevalent, we suspect that this topic will be a subject of debate in the near future and a part of community standards in the long run. These standards will shape how one searches for datasets and manages uncertainty throughout the trial.

Hypothesis Testing The main workhorse of Hephaestus is hypothesis testing. Here, the engine composes queries to the source databases to evaluate each proposed intervention. The VE partitions the samples into control blocks, and applies the accuracy metrics to each one. In the skin cancer example, the meta-system composes queries to extract the percentage incidence of cancer in the various test and control groups as demonstrated in Figure 1. By translating the accuracy metric into SQL queries. it extracts a hypothesis's score. The engine then evaluates all of the control blocks, computing a p-value for each.

A scoring function may be a single metric, like p-values, or a composite of several measures. For example, a researcher might want to test their discovery using p-values and cross-validation to see if their findings generalize cleanly to previously unseen data. They would then provide a function for combining the two metrics.

When evaluating each block, getting the controls right is crucial, and possibly one of the hardest parts of formulating a VE. There are some statistical techniques for inferring the null hypothesis under specific circumstances [28], but more work is needed in this area. As outlined in Section 2, control creation is a complicated part of experimental design.

Once the controls are established, the VE can start its test of the individual blocks. The user supplies an input and output schema to the analysis, as well as aliases to map from the source data to the VE. This analysis may be simple or sophisticated. Some experiments will execute basic aggregation, like the percentage in the oncology example. Hephaestus will be sufficiently general to permit users to pose more broad queries, using tools like Eureqa [48]. This system automatically attempts to fit a library of math formulas to a set of variables to predict an outcome. One could imagine starting their study with "fishing expedition" queries like this, followed by iteratively refining their search to more specific interventions. Hephaestus will make use of feature engineering [4] to infer relationships in these VEs.

After calculating the per-block statistics, the engine combines the results using well-known techniques [34]. This step calculates a weighted composite score for the hypothesis.

Intervention Ranking As Hephaestus evaluates a set of hypotheses, it accumulates and sorts them by score. To a first approximation, this is a generalization of top-k queries. On the other hand, optimizing this for VEs is hard owing to the presence of uncertainty and non-monotonic scoring functions. We discuss this further in Section 4.

3.2 Probabilistic Causal Graphs

VEs enable users to identify correlations that are statistically strong candidates for causal relationships. This property is important for users to determine *what* is present in their data. It does not, however, say anything about *why* the relationships exist. To do this, the researcher needs to look at her experiments in the context of a larger body of research. This will enable her to see the areas where the state of the art is consistent and spot anomalies as they arise.

Causal graphs [43], directed acyclic figures containing collections of cause-and-effect relationships, are well-suited to this goal. This symbolic language was developed to help re-

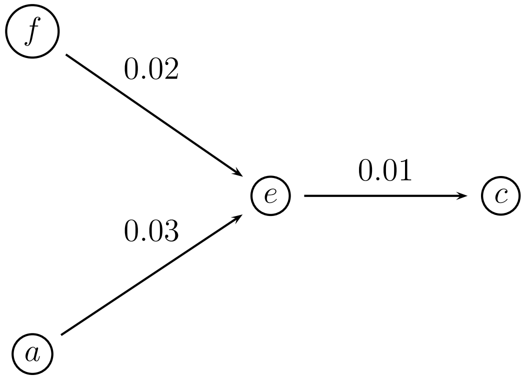


Figure 2: Probabilistic causal graph of skin cancer.

searchers integrate statistical methods with domain-specific knowledge. Hence, Hephaestus will maintain a directed acyclic graph denoting relationships that the experimenter has marked as potentially causal for analysis. Owing to the VE's use of statistical hypothesis testing, we envision this as a probabilistic causal graph (PCG) in implementation. To the best of our knowledge, this is the first proposal we have seen of this concept and data structure.

To continue the running example, after running her VEs, the oncologist revises her experiments to test whether fair skin and high-altitude living impact a test subject's overall sun exposure. The results of her VE are in Figure 2. She adds three hypotheses into a causal graph: fair skin (f), sun exposure (e), and high-altitude living (a). She demonstrates that all three are linked to skin cancer, c, although two of them impact it indirectly by modifying a subject's level of sun exposure.

PCGs differ from belief networks because they model conditional probabilities whereas causal graphs capture interventional probabilities. The former is learned using observational studies and it manages a graph's edges with Bayesian variables, and its relationships read like, "What is the probability of seeing skin cancer given the subject has experienced sun exposure?". Interventional probabilities model the system behavior if an intervention were directly applied, so it answers questions of the kind, "What is the probability this person will get cancer if we exposed them to the sun?". To date, no one has studied modeling these kind of graphs at scale. Below, we explore the definition of these graphs and the types of queries they are likely to service.

In the long term, VEs have the potential to empower community science efforts by assembling the experiments into large-scale PCGs. This would help specialized experts in related research areas leverage one another's work. Using this system, scientists would also be able to visualize their work in a larger context. It is easy to imagine such graphs having thousands of nodes and edges. In fact, a map of human cell signaling for cancer research has 1,600 nodes and 5,000 edges [13]. This graph was derived from the work of many biologists. Clearly, this data structure exceeds the expertise of a single person and requires groups of researchers working cooperatively to fully understand it. U.S. DARPA has identified causality modeling as a high priority for the advancement of research [53], although they consider it in the context of natural language processing and not statistical hypothesis testing for open science data.

Definition The nodes in a PCG will describe variables that may be an intervention or effect in any number of causal relationships. Graph edges denote suspected causal relationships, e.g., $A \implies B$. Each relationship has a weight associated with it, and this corresponds to the score of the edge's corresponding VE. This weight will be incrementally updated using continuous verification, reflecting the relationship's quality as new data accumulates. It is easy to imagine scientists subscribing to updates for relationships in their area of research.

When modeling a relationship containing greater than two nodes, a PCG may take one of three forms, as formalized in [23]. Some will have cascading relationships, where the effect is directly manipulated by one node that is influenced by a third variable. The relationship $f \implies e \implies c$ in Figure 2 is an example of this. Other nodes may exhibit a common cause, a 1:N relationship where a single node effects two or more variables. Lastly, a relationship may be described as common-effect, N:1, where a single node is jointly influenced by two or more interventions. Sun exposure is one such common effect variable, because it composes the impact of the other two interventions.

Composite relationships, containing an arbitrary number of nodes, can be partitioned into these primitives for evaluation. They enable the researcher to verify the accuracy of a science paradigm holistically. By identifying weak subgraphs, the platform will pick up on anomalies. This is crucial in eScience owing to the scale of the data and PCGs [12]. These areas of low-scoring VEs might imply a blind spot in the big picture or at least an area warranting further investigation. By assembling groups of anomalies, this will enable researchers to critically reason about the strength of a paradigm's underlying assumptions.

Computational Verification PCGs will let researchers delegate the complicated, tedious process of verifying the coherency of a paradigm to Hephaestus. These graphs may aid in the aforementioned search for weak points in a model. They will also be useful for identifying complementary strengths between rival science paradigms.

Assembling a PCG will enable researchers to identify competing explanatory relationships within a single graph. For example, if two VEs from the literature have an effect of A, the engine can rapidly identify this and score each hypothesis, determining whether it is likely a joint effect or conflicting observations. It could also present the pair to the user who might then decide how to resolve it.

A graph verifier will test whether published findings from many studies all fit within a globally consistent set of relationships. This structure may be verified algorithmically. The checker will test for invalid conditions, such as causal loops, such as $A \implies B \implies C \implies A$. It will also monitor rules provided by domain experts. Fast algorithms already exist for many of these problems [51].

A computational verifier will also need to test for inconsistencies in the results of VEs. Simpson's Paradox, as discussed in Section 2.1, is one such issue. This and other statistical anomalies are easy to find computationally, but difficult for humans to detect.

Hephaestus may also support user exploration of subgraphs from one or more PCGs. Hence, if they are comparing two science paradigms, they might ask the meta-system to identify matching subsets of their graphs where one outperforms the other. They may also search for contradictions between the two graphs if they contain intersecting nodes.

This verification will also extend to continuous verification of VEs. As more data becomes available, Hephaestus will recompute the weights of its edges by applying the VEs to the new sources. This may promote some relationships over others in the case of competing VE-backed theories. It also has the possibility of breaking some of the graph relationships, and this may create ripples of change throughout the structure. Hence, the platform will need techniques to efficiently test the integrity of a graph incrementally, rather than recompute the whole set.

Exploratory Analysis PCGs will enable man-machine symbiosis by helping researchers explore a hypothesis space. They will do so by visualizing proposed relationships, by adjusting their model parameters, and by interactively probing the graph.

A graph will render proposed relationships from new VEs over the findings that are already accepted. This will be especially useful for VEs containing the wildcard. The new ones might appear as dotted lines on top of the existing nodes and edges so that the scientist can assess how they would fit into the larger picture. The user could the query the proposed edges to see any impact they might have on the consistency of the graph. They could optionally accept the ones they deemed most interesting to their personal collection of theories.

The user may also dynamically adjust parts of their VE design, seeing how it effects the graph as a whole. The PCG visualization might come with sliders for setting the threshold of a hypothesis test, where increasing or decreasing the tolerances of the VE would make graph edges appear or disappear accordingly. They could also manipulate the scoring function and other parts of the experimental design.

Researchers may zoom in and out of the graph at different levels of abstraction. If a scientist is looking at how clusters of cells interact with one another when cancer cells grow, they might zoom in to see the relationships between organelles in a single cell. Each level will contain reference to different bodies of work, and mean plugging in to the graphs of researchers in neighboring fields.

The graph will also facilitate visualizing the strength of its relationships in different subgraphs. If a causal link breaks owing to continuous verification of the underlying VEs, the platform will alert the user to new anomalies. Users could also generate a heat map, showing where the graph has the strongest statistical significance and where the model is strained. Clearly, there are huge gains possible by maintaining this network of suspected causal relationships and sharing them with others. By formalizing bodies of research as graphs, researchers will computationally verify the assumptions that underpin their work and visualize how new results fit into the prior work.

4. RESEARCH CHALLENGES

The vision of Hephaestus gives rise to several important challenges for the data management community. Many are interdisciplinary and are well-positioned for collaborations with statisticians and human-computer interaction researchers. In this section, we explore the implications of integrating statistics for hypothesis testing into query optimization. We then look at the challenges associated with man-machine symbiosis for VEs and PCGs. Lastly, this section contains an outline of several research opportunities for the architecture of this meta-system.

4.1 Integrated Statistics

Although we propose Hephaestus as a meta-system on top of existing databases, it will benefit from working with storage engines optimized for statistical analysis. Rather than decoupling the query processing on open data repositories from statistical hypothesis testing, this issue lends itself to an integrated approach as demonstrated by BlinkDB [2]. This database provides approximate query results over large datasets using sampling. It taps into statistics about the underlying data's distribution to compute results with bounded errors.

Hephaestus may benefit from a similar approach where it takes into account the source data distribution and operator characteristics to compile VEs into relational-style query plans. Below we detail several statistics challenge that are amenable to performance optimization by integrating them into the query optimization process.

Combining Disparate Datasets At present, putting multiple datasets into a single analysis takes careful manual planning. Limited techniques exist for this issue [7], and the available options are targeted for data sources that share a schema. There is work to be done in aggregating over the results of many studies with differing degrees of overlap in their experimental design. A platform may improve query performance by selecting the most efficient intermediate representation for each study and ordering the computation of each result to rapidly rule out hypotheses.

The first version of Hephaestus will calculate the accuracy metrics of each dataset independently, and take a weighted sum over the control blocks shared among studies for the hypothesis's score. The experiment designer will select a weighting function; they are likely to use factors such as each block's sample size or variance. This approach is appealing because if the datasets are not collocated, it simplifies query planning by not aggregating all of the samples into one mega-study.

Missing Data Imputation One aspect of data reuse that makes designing experiments non-trivial is that not all of the variables the experimenter wants to account for may be present in every dataset. To address this, the researcher may reduce their control blocks to the intersection of the source schemas. They can then analyze the variance of their experiment blocks to see if this is satisfactory.

If these reduced sets of controls are insufficient, the researcher may statistically infer the missing variables from more complete samples [20]. Any uncertainty introduced from this process needs to be propagated through the rest of the analysis. This typically involves applying a modeling function, like linear regression, to the data in order to learn from the fully populated dataset the likely values of the unavailable variables. Identifying opportunities to apply imputation at scale and optimizing this process over distributed, heterogeneous data sources is an open problem. The same is true about efficiently managing the uncertainty created by this technique for massive datasets.

False Discovery Rate Support Data-intensive science has the novel potential that researchers can test an unprecedented number of hypotheses over a single dataset. For example, a genomics study with n human subjects may record thousands of measurements per person. It is likely that correlations will emerge from this analysis that pass the threshold of statistical hypothesis testing, but are nonetheless are spurious because the number of potential hypotheses vastly exceed the count of human subjects. There are a variety of techniques for controlling the number of false positives [6], such as taking some fraction of all of the hypotheses that tested true. This fraction is selected from the hypotheses that have the highest accuracy ratings.

At first glance, this problem is similar to a top-k selection, but the k is not known up front because it is a function of how many hypotheses make it over the bar. Moreover, k only grows with time as more hypotheses pass the threshold. Optimizing this adaptive cutoff of the query results is an unsolved challenge.

Sampling Another possibility for query speedup is to use sampling to estimate the hypothesis with high confidence over a smaller subset of the data. If the user is willing to accept some uncertainty, perhaps making the results within 95% of the correct figure, many scoring functions are amenable to sampling.

As demonstrated by BlinkDB, integrating principled sampling into query execution dramatically speeds up the process. There are several generalizations needed to this framework to make it applicable for data reuse. This database leverages precomputed sample sets for its fast, approximate answers. For data that is not stored locally, this approach may need to create composite samples from multiple sources, each of which will be of varying size and may have different data distributions.

A related issue arises when the samples used for approximate query processing produce inaccurate error bounds. The research in [1] demonstrated that techniques for deriving error bars on approximate relational query results produce high error rates in practice. The authors created ways to estimate these errors and use them to either enlarge the error bars or to report that sampling is not possible. Clearly more work is needed for adaptive, iterative sampling within the query planner. This class of queries will need error bars and primitives describing their quality as first-class objects in its evaluations in order to converge on satisfactory solutions quickly.

4.2 Man-Machine Symbiosis

We now examine a set of open challenges regarding how to efficiently use human attention to accelerate scientific discovery. These problems revolve around making the computation of results fast enough for interactive visualization. We also look at the conditions under which the engine will need to alert the user to ambiguities in the datasets and hypothesis testing results. We also briefly touch on the issue of empowering researchers to prune the space of hypotheses and visualize uncertainty.

Incremental Graph Evaluation Probabilistic causal graphs call for a rich set of interactions, as outlined in Section 3.2. First off, the visualizer will need techniques to store the results from VEs in a way such that they can be combined with new runs of the same VE over different input data. Having these intermediate results would enable the database to combine the hypothesis tests from multiple datasets into a single scoring function without rerunning the previous VEs. Materialized views may provide an efficient way to incrementally compute this figure for one or more VEs, but they may require new building blocks in order to support complex scoring functions. Second, the PCG engine would benefit from working with sampling as outlined above in order to rapidly recompute the visualization when the user modifies parameters such as their hypothesis testing threshold.

Hypothesis Space Modeling One approach to taming the complexity of mining a large number of hypothesis is to selectively tap into human intelligence. When a VE is proposing interventions owing to a wildcard operator, the engine could display a partial list to the experiment writer and ask them to eliminate ones they deem uninteresting or irrelevant. If a community of users leverages Hephaestus, it may be possible to learn from this feedback collectively. Hence, domain experts could provide rules like "the weather is never affected by a person's pulse", and anyone can use them. Ultimately, we suspect that a hypothesis space will be pruned using a combination of sampling, feature engineering, and crowdsourcing.

A second challenging aspect of hypothesis space modeling in Hephaestus is handling complex correlations. If the experiment designer asks a question of the form $A \implies B \implies C$, they want to score each link in the chain. If A and B are both a large set of proposed interventions, this will create an explosion in the space of correlations to quantify. Clearly this won't scale up. VEs where many interventions contribute to a single outcome will also call for sophisticated modeling to select the most plausible hypotheses for user feedback.

Another scenario where this hypothesis space could get complicated to model is for conditional interventions. Some causal relationships are not simply stated with "A implies B", hence Hephaestus may need to create forks in its causal graph. In the running example, the VE may determine that the youngest cohort in the study has high-altitude living as its strongest intervention, whereas the elder control block's cancer rate is more influenced by skin tone.

Language Design VE language design involves understanding and meeting the requirements of eScientists expressed in a form that is amenable to database optimization. Examples include expressing empirical requirements such as multiple, ordered hypotheses with scoring and blocking and allowing subqueries in a VE. We introduced a simple language in Section 3.1, but clearly a richer model is needed to express VEs. In particular, if nested queries have wildcards, careful thought is needed to find efficient ways to create relational-style plans. This is another place where feature engineering is likely to make VEs more efficient.

Managing Inconsistency Building in mechanisms for inconsistency in eScience will be necessary make VEs effective. Presently there is no principled way to differentiate complex interactions from confounders and human experts are needed to intervene for these circumstances. Instead, we will focus on identifying inconsistencies and finding ways to economize the user's time for the ones that are most likely to yield results. The system also needs to be flexible for a diverse set of directives in response to these alerts. Expert feedback might include "eliminate this intervention", "the order is most probably this, and these ones seem plausible", or " evaluate these combinations".

Visualizing Uncertainty Another interesting challenge in this framework is presenting uncertainty to the user. Although error bars are needed for nearly every step in Hephaestus, displaying them on a graph is not a solved problem. There are many possible ways to do this, such as varying the thickness of the edges and color coding relationships by their VE's score. User studies are needed on the best way to convey this important element of the experimental results.

4.3 Architecture & Performance

There are several challenges associated with building this platform to make it both accurate and performant. Here, we outline a few of the architectural questions. Many of them are likely to build on existing database research, and we sketch out these approaches when applicable.

Source Data Search As we touched upon earlier, identifying the best data sources for a VE will call for a mix of conventional search techniques and specialized ones to accommodate the needs of statistical hypothesis testing. In particular, users will want to take advantage of metadata about the provenance and schema of potential data sources. Work will need to be done to identify ones that meet the experiment's design, even in the presence of data that may vary from empirical to abstract in the context of a VE.

Query Translation Once a language is researched and established for VEs, the Hephaestus engine will need to be able to compile it into queries for the open data repository. These queries might be in SQL or any number of domain-specific languages depending on how the data is stored. Finding the right building blocks for rewriting VEs for one or more storage engines is an open question.

Aggregation VEs with many possible interventions are likely to benefit from reframing multiple hypothesis tests as data cube queries. Rather than executing the query in Figure 1 once per intervention, Hephaestus could compose and evaluate bins from multiple hypotheses at the same time. If the system groups by the proposed interventions, the binning query in Figure 1 becomes `SELECT count(*) FROM cancerSubjects GROUP BY age, gender, intervention1, intervention2, ... CUBE(<interventions>)`. This would aggregate the count at every level, potentially reusing intersecting sets of controls. This batching would also be useful for probabilistic causal graph verification for ambiguities like Simpson's Paradox. Identifying the right levels of aggregation and coordinating this effort among multiple data sources is an open question.

Uncertainty The database community has created a variety of methods for managing uncertainty in relational data [14, 18, 54]. It is unclear how to propagate the models for these solutions through the steps of a VE, especially in the presence of multiple evaluations of the same VE over different data sources, and in cascading relationships. More work is needed to find efficient ways to complete these queries.

Top-k Generalization If the VE limits the number of results returned, this may reduce the search space of hypotheses to evaluate. Taking a page from relational top-k optimization, there are at least two vectors for this approach. First, the database would use sampling to evaluate thresholds, such as the p-value being ≤ 0.05. Interventions that do not meet this requirement would not compute any additional metrics for complex relationships.

Another direction is to opportunistically ordering the evaluation of multiple datasets for a single intervention. Here, if the results are combined by a weighted sum on their sample size a la [7], the optimizer can determine the bounds of the smaller dataset's scoring function that will result in the intervention being rejected. This would prevent a second round of testing on the larger data source. There are several aspects to the VE's structure that make this type of optimization challenging. Supporting false discovery rates, uncertainty, and complex analysis like p-values makes it hard to model the outcomes of different hypotheses. Sampling will be key to speeding up this process.

Distributed Query Optimization A related issue to top-k generalization is that of coordinating queries over multiple data sources. This work can be made more efficient by taking into account the relative capabilities of each member database to select the storage engine that will run first. Running the smallest datasets will efficiently eliminate hypotheses. In addition, if data needs to be moved from one host to another for joins and other comparisons, the optimizer will need to accurately estimate the cost of these operations and availability of hardware resources on each host.

In summary, we outline the first research steps needed to create a data reuse platform. We examine the integration of statistics, optimization of man-machine symbiosis, and a set of architecture challenges associated with open science data management. Each has the opportunity to build on existing database solutions, but still has numerous novel directions.

5. CONCLUSIONS

In this vision paper, we explore how science is changing as research data becomes more abundant and open. We note that this creates an opportunity to statistically test hypotheses on existing data in many circumstances. To this end, we propose Hephaestus, a platform for data reuse in eScience. This system will enable scientists to explore their theories in two ways. First, virtual experiments (VEs) are designed for statistical hypothesis testing from empirical in-house trials, publicly available open science repositories, or a combination of the two. These queries will simulate randomized controlled trials by implementing the principles of empirical scientific research. VEs will estimate the statistical significance of correlations using a scoring function supplied by the user. We then propose to assemble the correlations found by Hephaestus and by existing empirically-derived scientific discovery into probabilistic causal graphs, so that researchers can share and inspect their findings, updating them as discoveries are made. This framework will let researchers create experimental designs that are testable on any number of data sources. Hence, they have the opportunity to perform continuous verification on their discoveries as new data becomes available from related research.

This work puts forth numerous research directions for the data management community. In particular, we advocate for integrating statistics more closely with query execution in science databases, carefully rationing human attention for hypothesis selection, and generalizations to relational-style database architecture to support VEs. This work is a first step toward enabling eScience practitioners to mine reality from massive datasets for scientific discovery.

6. ACKNOWLEDGMENTS

The authors thank their colleagues for their insightful feedback, including Jaime Carbonell, Herb Lin, Sam Madden, Gregory Piatetsky-Shapiro, Michael Stonebraker, Eli Upfal, and Eugene Wu. We are also grateful to the anonymous reviewers of this paper whose comments guided many important improvements. We also acknowledge the many scientists who have so generously taken the time to answer our questions on how scientific research is conducted, including Clare Bernard, James Michaelson, Tanya Monro, Pardis Sabeti, and Pete Szolovits. This work was funded by the Intel Science and Technology Center for Big Data.

7. REFERENCES

[1] S. Agarwal, H. Milner, A. Kleiner, A. Talwalkar, M. Jordan, S. Madden, B. Mozafari, and I. Stoica. Knowing when you're wrong: Building fast and reliable approximate query processing systems. SIGMOD '14, pages 481–492, 2014.

[2] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica. Blinkdb: queries with bounded errors and bounded response times on very large data. In *Eurosys*, pages 29–42. ACM, 2013.

[3] C. Anderson. The end of theory: the data deluge makes the scientific method obsolete. *Wired*, 14(6), 2008.

[4] M. Anderson, D. Antenucci, V. Bittorf, M. Burgess, M. J. Cafarella, A. Kumar, F. Niu, Y. Park, C. Ré, and C. Zhang. Brainwash: A data system for feature engineering. In *CIDR*, 2013.

[5] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

[6] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[7] M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein. *Introduction to meta-analysis.* Wiley, 2011.

[8] P. G. Brown. Overview of scidb: large scale array storage, processing and analysis. In *SIGMOD*, pages 963–968, 2010.

[9] C. Carilli and S. Rawlings. Science with the Square Kilometer Array: motivation, key science projects, standards and assumptions. *arXiv preprint astro-ph/0409274*, 2004.

[10] CERN. Large hadron collider. `http://home.web.cern.ch/topics/large-hadron-collider`.

[11] C. Charig, D. Webb, S. Payne, and J. Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British medical journal (Clinical research ed.)*, 292(6524):879, 1986.

[12] Committee on the Analysis of Massive Data. *Frontiers in Massive Data Analysis.* The National Academies Press, 2013.

[13] Q. Cui, Y. Ma, M. Jaramillo, H. Bari, A. Awan, S. Yang, S. Zhang, L. Liu, M. Lu, M. O'Connor-McCourt, et al. A map of human cancer signaling. *Molecular systems biology*, 3(1), 2007.

[14] N. Dalvi, C. Ré, and D. Suciu. Probabilistic databases: diamonds in the dirt. *CACM*, 52(7):86–94, 2009.

[15] P. A. David. The economic logic of "open science" and the balance between private property rights and the public domain in scientific data and information: a primer. In *The role of scientific and technical data and information in the public domain*, pages 19–34. Basic Books, 2003.

[16] D. De Roure, C. Goble, and R. Stevens. The design and realisation of the virtual research environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5):561–567, 2009.

[17] A. R. Diekema, A. Wesolek, and C. D. Walters. The NSF/NIH Effect: Surveying the Effect of Data Management Requirements on Faculty, Sponsored Programs, and Institutional Repositories. *The Journal of Academic Librarianship*, 2014.

[18] A. Faradjian, J. Gehrke, and P. Bonnet. GADT: A probability space ADT for representing and querying the physical world. In *ICDE*, pages 201–211. IEEE, 2002.

[19] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

[20] J. S. Greenlees, W. S. Reece, and K. D. Zieschang. Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77(378):251–261, 1982.

[21] F. Griffiths and K. Jones. The use of hormone replacement therapy; results of a community survey. *Family practice*, 12(2):163–165, 1995.

[22] W. Gunn. Reproducibility: Fraud is not the big problem. *Nature*, 505(7484):483–483, 2014.

[23] Y. Hagmayer, S. A. Sloman, D. A. Lagnado, and M. R. Waldmann. Causal reasoning through intervention. *Causal learning: Psychology, philosophy, and computation*, pages 86–100, 2007.

[24] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Sys.*, 24(2):8–12, 2009.

[25] D. Halperin, V. T. de Almeida, L. L. Choo, S. Chu, P. Koutris, D. Moritz, J. Ortiz, V. Ruamviboonsuk, J. Wang, A. Whitaker, S. Xu, M. Balazinska, B. Howe, and D. Suciu. Demonstration of the Myria Big Data Management Service. In *SIGMOD*, pages 881–884, 2014.

[26] I. F. Ilyas, G. Beskales, and M. A. Soliman. A survey of top-k query processing techniques in relational database systems. *Computing Surveys*, 40(4):11, 2008.

[27] J. P. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[28] V. E. Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48):19313–19317, 2013.

[29] T. S. Kuhn. *The structure of scientific revolutions.* University of Chicago press, 1962.

[30] A. Kumar, F. Niu, and C. Ré. Hazy: Making it easier to build and maintain big-data analytics. *CACM*, 56(3):40–49, 2013.

[31] D. A. Lawlor, G. D. Smith, and S. Ebrahim. Commentary: The hormone replacement–coronary heart disease conundrum: is this the death of observational epidemiology? *International Journal of Epidemiology*, 33(3):464–467, 2004.

[32] D. Lazer. Mistaken analysis. *MIT Tech Review*, 2014.

[33] J. C. R. Licklider. Man-computer symbiosis. *Human Factors in Electronics, IRE Transactions on*, (1):4–11, 1960.

[34] T. Liptak. On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, 3:171–197, 1958.

[35] S. Lohr. Google flu trends: the limits of big data. *New York Times*, 2014.

[36] G. Marcus and E. Davis. Eight (No, Nine!) Problems With Big Data. *New York Times*.

[37] G. A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[38] NASA. Moderate-Resolution Imaging Spectroradiometer (MODIS). 1999. `http://modis.gsfc.nasa.gov/`.

[39] National Institutes of Health. Big data to knowledge (bd2k) initiative. `http://bd2k.nih.gov/`.

[40] National Institutes of Health. Biomedical information science and technology initiative. `http://www.bisti.nih.gov`.

[41] National Institutes of Health. NIH Grants Policy Statement. `http://grants.nih.gov/grants/policy/nihgps_2013/`.

[42] F. Niu, C. Zhang, C. Ré, and J. W. Shavlik. DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. In *VLDB*, pages 25–28, 2012.

[43] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[44] J. Pearl. *Causality: Models, Reasoning and Inference*, volume 29. Cambridge Univ Press, 2000.

[45] J. Robertson. Stats: We're doing it wrong, Apr. 2011.

[46] M. Rubacha, A. K. Rattan, and S. C. Hosselet. A review of electronic laboratory notebooks available in the market today. *Journal of the Association for Laboratory Automation*, 16(1):90–98, 2011.

[47] N. Savage. Automating scientific discovery. *CACM*, 55(5):9–11, May 2012.

[48] M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[49] N. Silver. *The signal and the noise: Why so many predictions fail-but some don't*. Penguin, 2012.

[50] S. Spangler, A. D. Wilkins, B. J. Bachman, M. Nagarajan, T. Dayaram, P. Haas, S. Regenbogen, C. R. Pickering, A. Comer, J. N. Myers, et al. Automated hypothesis generation based on mining scientific literature. In *SIGKDD*, pages 1877–1886. ACM, 2014.

[51] R. E. Tarjan and M. Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on computing*, 13(3):566–579, 1984.

[52] J. A. Tyson. Large synoptic survey telescope: overview. In *Astronomical Telescopes and Instrumentation*, pages 10–20. International Society for Optics and Photonics, 2002.

[53] U.S. Defense Advanced Research Projects Agency. Big mechanism seeks the "whys" hidden in big data. 2014. `http://www.darpa.mil/NewsEvents/Releases/2014/02/20.aspx`.

[54] J. Widom. Trio: A system for integrated management of data, accuracy, and lineage. *Technical Report*, 2004.